Early experience with Montecito

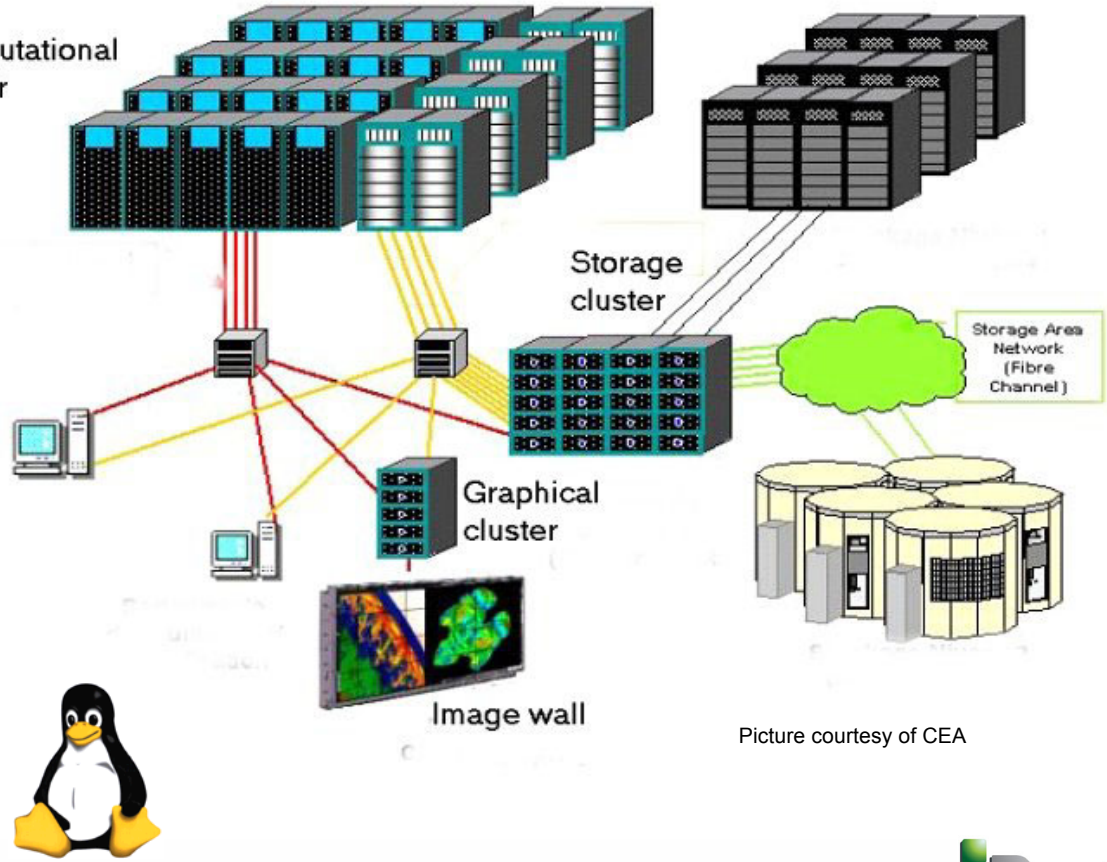
Daresbury December 6th

Bull R&D Echirolles centre: Denis Foueillassar

Cluster environment and partners

Customer

Project partners



Picture courtesy of CEA

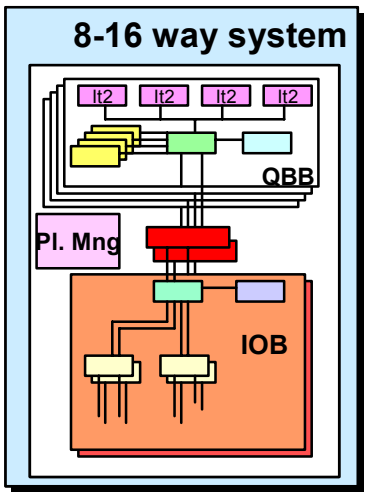
Cluster configuration

- Performance objective — > 50 Teraflops (Linpack)
- Compute nodes — 544 NovaScale nodes (~200 racks)
- Total memory — 30 Tbytes
- Disk space — 1 Pbytes - 56 nodes (54 OSS+2 MDS)
- Packaging — 3 nodes per rack
  16 cores per node

Why we won this bid

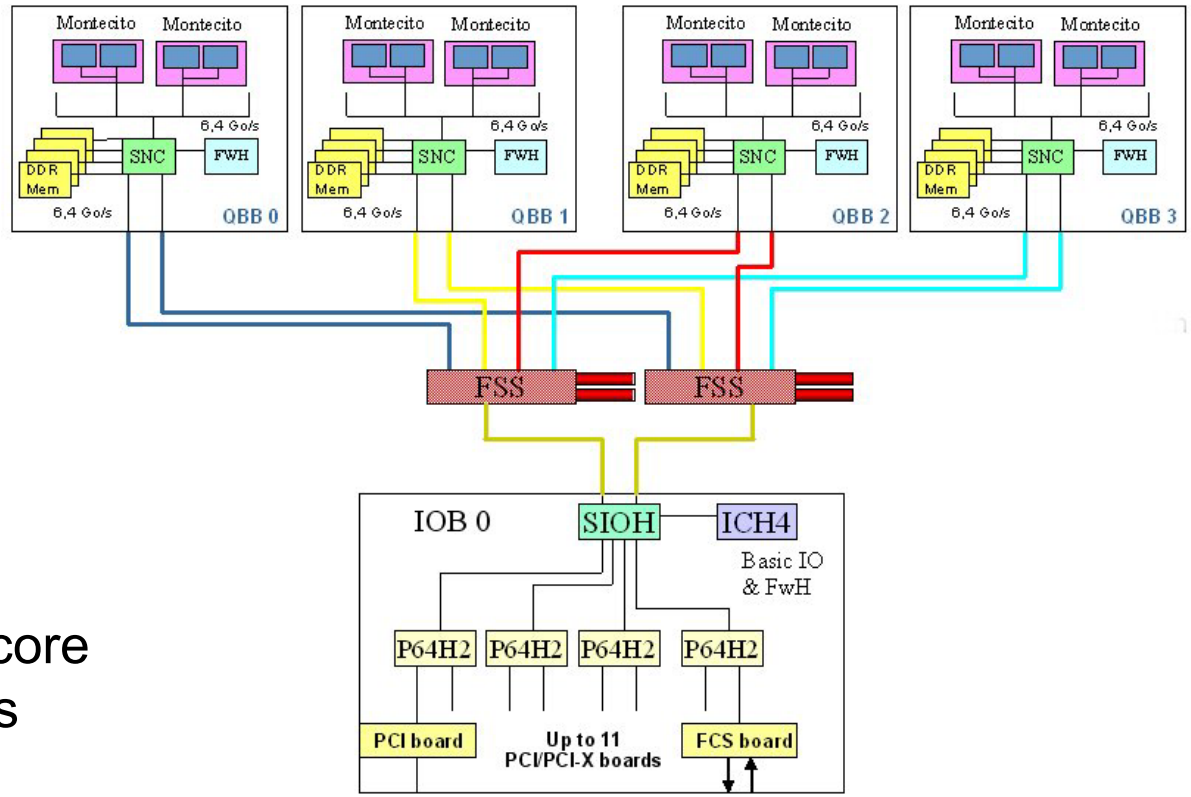
- Performance :
  - Itanium2 floating point calculation.
  - Interconnect throughput and latency.
  - Input/Output throughput.
- Standards and Open Source software for OS and management tools.
- Strategic partnership.
- Global cost of ownership over four years.
- Local Team competency
  (services & technical support).

Montecito vs. Madison first approach



Madison
QBB

Montecito

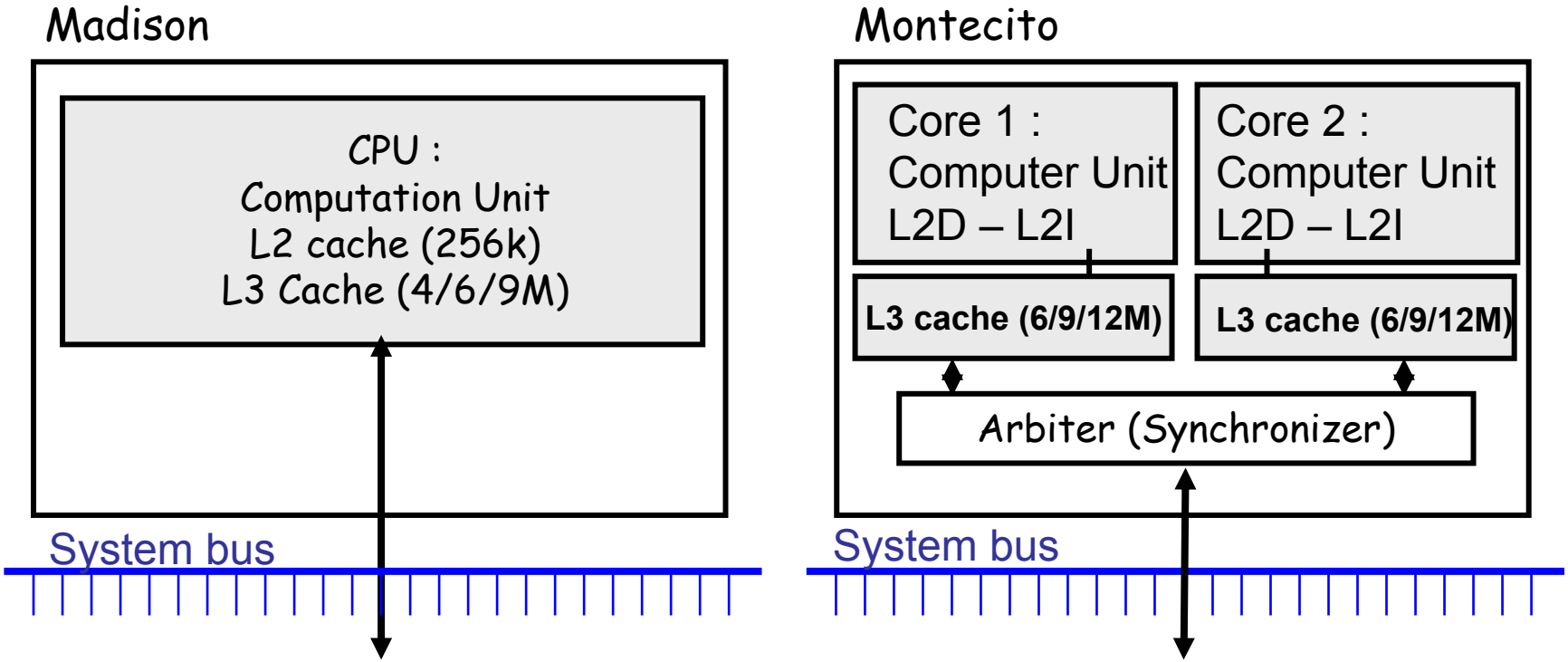


Both configurations
- frequency 1.6 GHz
- cache size 9 Mbytes / core
- memory size 48 Gbytes

Observations with 16 cores on a node

- On a DGEMM measurement (Dual precision Matrix Multiply) with Intel MKL the measure gives 98.5 GFlops compared to peak value of 102.4 GFlops (6.4 GFlops x16)
  - more than 94% efficient
- On a HPL Linpack a node was able to deliver 92.5 GFlops compared to peak value 102.4
  - about 90% efficient.
- Other benchmarks demonstrated a comparable global throughput improvement.
- At the same frequency and same cache size and using different benchmarks, a Montecito core performance is equivalent to a Madison processor

Differences Madison / Montecito



Montecito L2 cache splits Data and Instruction
(256K Data + 1MB Instruction instead of 256K global)

One memory bus cycle lost in bus arbitration between the 2 cores at each change of data sender : This increase lightly the memory access latency , but does not impact the bandwidth.

Montecito: Kernel adaptations

- Kernel change
  - To support SAL/PAL interface (threads/cores topology).
  - To support system call mprotect (L2 split Instruction/Data).
  - Logical processor numbering changed to be in conformance with Bull hardware specification.
- Scheduler change
  - To support new hyperthreading domain (2 threads for one core).
    - Default initial placement has to use one thread per core.
    - Idle thread releases core resources for the other thread.
- Logical partitioning change
  - “cpuset” changed to support one or two threads per core.
- Loader change
  - X Windows Server (L2 split).

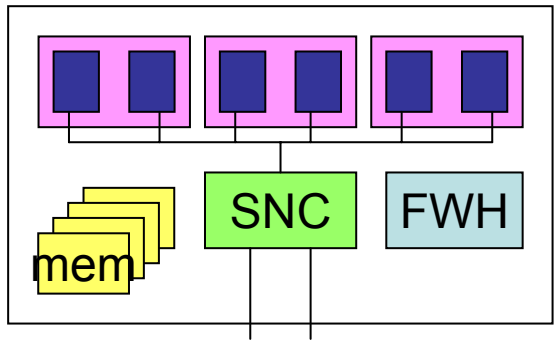
Montecito:
Intel compiler and libraries adaptations

- Compilers
  - The code generated for Madison runs the same way on Montecito.
  - An extension in the compiler allows to generate specific Montecito cases (mainly the additional shift unit)
  - The Itanium2 instruction set is very powerful
  - A correct generation of code with adequate prefetch is better.
  - An OpenMP bug has been corrected by Intel compiler team.
- Intel tools
  - Vtune very useful
- MKL libraries
  - Are mandatory to get the best performances of the Itanium2. Intel has helped us to make some customizations related to the NUMA aspects of our machines.

Upgrade test in the lab.

- Power consumption per Montecito socket (Less than 130 Watts) is quite similar to Madison processor.
- Heat dissipation follows the same curve,
  - air conditioning can be decreased for the same power.

- Montecito is half the size of Madison, to take benefit of this we did some tests with three Montecito sockets per QBB (6 cores per QBB and 24 cores per system). Performance per node is increased by almost 50%



Observations with 24 cores per node

- On a DGEMM measurement (Dual precision Matrix Multiply) with Intel MKL the measure gives 145 GFlops compared to peak value 153.6 GFlops.
  - more than 94% efficient
- On a HPL Linpack with small memory (still 48GB, but divided by 24 process instead of 16) a node was able to deliver 129.5 GFlops
  - about 85% efficient
  - This performance can be improved increasing the memory size.

Conclusion

- Montecito is performant and socket price is (or will) be approximately the same as Madison
  Consequently price performance is doubled.
- Hyperthreading is also an area of possible performance improvement - more time needed.
- Good cooperation with Intel gives us confidence to reach the 50 Teraflops Linpack at CEA.
- Waiting for GA of Montecito to ship Montecito in Bull NovaScale to other customers.

Architect of an Open World™